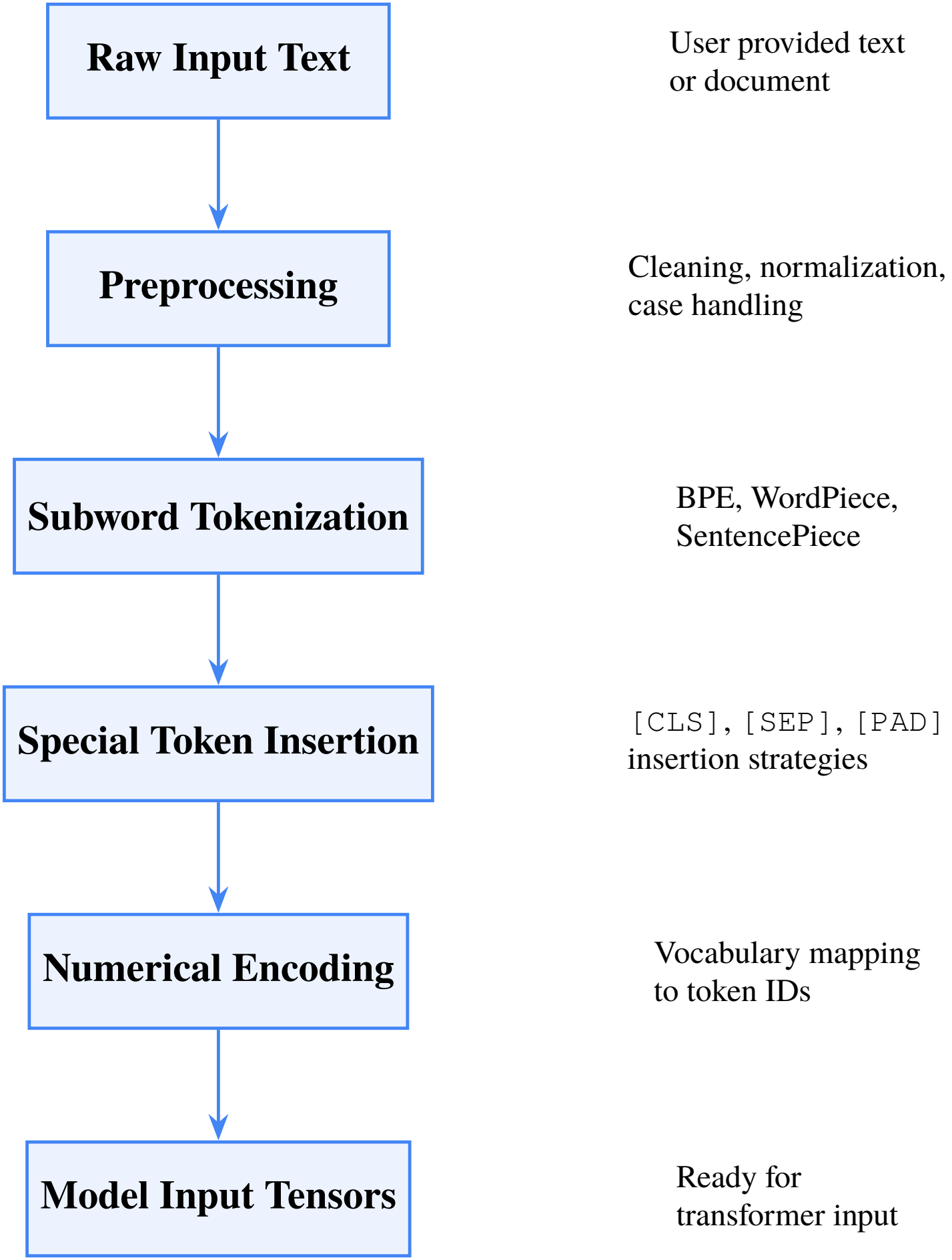
Raw Input Text
User provided text or document
Preprocessing
Cleaning, normalization, case handling
Subword Tokenization
BPE, WordPiece, SentencePiece
Special Token Insertion
[CLS], [SEP], [PAD] insertion strategies
Numerical Encoding
Vocabulary mapping to token IDs
Model Input Tensors
Ready for transformer input